DOCUMENT RESUME

ED 170 381 TM 005 237

AUTHOR Berk, Ronald A.
TITLE Some Guidelines for Determining the Length of Objectives-Based Criterion-Referenced Tests.
PUB DATE Apr 79
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavioral Objectives; *Course Objectives; *Criterion Referenced Tests; *Decision Making; Educational Objectives; Mastery Tests; Primary Education; Student Testing; *Test Construction; Test Items; Test Results
IDENTIFIERS *Test Length

ABSTRACT

Four factors essential to determining how many items should be constructed or sampled for a set of objectives are examined: (1) importance and type of decisions to be made with the results; (2) importance and emphases assigned to the instructional and behavioral objectives; (3) number of objectives; (4) practical constraints, such as item writing time, time available to administer the test, and student characteristics. Within the context of related research on cut-off scores and reliability, it is recommended that between 5 and 10 items per objective be employed for most classroom decisions and between 10 and 20 items be used for school, system, and state level decisions. Specific guidelines for teachers and evaluators, and an illustrative application using objectives from a second grade dictionary skills unit are provided. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED170381

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SOME GUIDELINES FOR DETERMINING THE LENGTH OF OBJECTIVES-BASED CRITERION-REFERENCED TESTS

Ronald A. Berk

The Johns Hopkins University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald A. Berk

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at the annual meeting of the National Council on Measurement in Education

San Franciso, April 1979.

TM009 237

Abstract

Four factors essential to determining how many items should be constructed or sampled for a set of objectives were examined: (1) importance and type of decisions to be made with the results; (2) importance and emphases assigned to the objectives; (3) number of objectives; (4) practical constraints. Within the context of related research on cut-off scores and reliability, it was recommended that between five and 10 items per objective be employed for most classroom decisions and between 10 and 20 items be used for school, system, and state level decisions. Specific guidelines teachers and evaluators can use and an illustrative application are provided.

ED 023 [illegible]

Some Guidelines for Determining the Length of Objectives-Based Criterion-Referenced Tests

The determination of how many items should be constructed or sampled for a set of objectives is one of the most perplexing issues in criterion-referenced measurement. There is no single answer to the question. The literature that has addressed the topic directly (Berk, 1978 ; Eignor & Hambleton, 1979; Flaner, 1974; Hambleton, Hutton & Swaminathan, 1976; Millman, 1973; Novick & Lewis, 1974; Wilcox, 1976, in press), or indirectly (Berk, 1976; Huynh, 1976; Subkoviak, 1976,1978) has concentrated primarily on the relationship of test length to test validity, ie., mastery - nonmastery classification decision accuracy. Few of the studies link the problem to test reliability.

In addition to the empirical findings to date, there seems to be a multiplicity of other factors the test-maker must consider in assessing test length. These include the following: (1) importance and type of decisions to be made with the results; (2) importance and emphases assigned to the objectives of the program; (3) number of objectives; (4) practical constraints, e.g., time anticipated for item writing and test administration. This paper will examine these factors and provide some guidelines teachers and evaluators can use in developing tests for placement, formative, and summative evaluation. The presentation will be couched within the context of the research cited above. An application of the guidelines to an illustrative set of objectives for a second-grade language arts unit will also be described.

Objectives-Based Scheme

Several strategies for specifying content domains and generating items have been proposed in recent years as alternatives to the objectives-based approach. They are IOX test specifications, item transformations, item forms, algorithms, and mapping sentences. Despite the rigor and precision with which these strategies can define a content domain and produce items to measure achievement of the domain, instructional and behavioral objectives have remained the predominant mechanism for building criterion-referenced tests.

While a variety of labels have been assigned to the different levels of educational objectives (e.g. ultimate, terminal, enabling), "instructional" and "behavioral" seem to be the most widely used and most meaningful terms. The former focuses on what a student is expected to know at the conclusion of an instructional program, e.g., unit, course, grade level; the latter defines the instructional objective in specific observable behaviors. It is the instructional objective that typically constitutes the level at which decisions about individuals and programs are made. This bilevel framework will furnish the foundation for determining the number of items.

The problem of deciding on the number of items is also a function of how the item domain is conceptualized. Either the specific number of items to be included on the test is written from the objectives and conceived as a sample from the domain of all possible items that could be written or an exhaustive finite domain of items is constructed and then sampled using a random or stratified random procedure. Although the explication of factors affecting test length apply equally to the questions of "How many items should be written for each objective?" and " How many items should be sampled?", the perspective in this paper is the former since teachers and evaluators rarely can draw item samples, even though many commercial test-makers have recently hastened to generate item-banks.

Four Factors

Type of Decisions

In general, the larger the number of (quality) items, the greater confidence one can place in evaluation decisions. The tentativeness or permanency of the decisions, however, in conjunction with the resultant frequency of Type I and II decision errors are primary considerations in arriving at the number of items essential to adequately measure an instructional objective. Classroom formative and placement decisions are often tentative whereas classroom summative evaluation, grade level or course placement, and program evaluation decisions are more permanent. It is recommended that between five and 10 items per objective be

employed for most classroom decisions and between 10 and 20 items be used for school, system, and state level decisions. The rationale for these recommendations is based on the availability of corroborative information on decision accuracy, the possibility and practicality of decision reversals, and the research on cut-off scores and reliability pertinent to test length. This rationale is discussed next.

Tentative decisions. The minimum number of items that should be used for individual decision-making is five. This number is a function of the frequency of errors that may result from placement and formative decisions and the tentative nature of those decisions. One type of decision error (Type II error or false mastery) occurs as a student is classified a master when, in fact, he/she has not mastered an objective. The error becomes evident as the student displays frustration with a new and often more difficult assignment or the student refuses to participate in the activities of the group into which he/she was placed. The teacher can easily reverse the original decision by assigning appropriate materials to remediate the student's skill weaknesses or by placing the student in another work group. The second type of decision error (Type I error or false nonmastery) relates to the converse situation where a master is mistakenly classified as a nonmaster. It can be resolved in a similar manner. The immediate feedback a teacher receives on placement or formative decision accuracy precludes the need to employ as many as 15 items to assess attainment of a single objective. At the other extreme it would not be desirable to use so few items (one or two) that close to half the decisions made would be in error. The research on test length seems to suggest that objectives consisting of between five and 10 items provide sufficient accuracy for most classroom decisions, (cf. Novick & Lewis, 1974).

Permanent decisions. In cases where there is considerable weight attached to the decisions and/or decision reversals are impossible or impractical, a minimum of 10 items per instructional objective should be used. Decisions regarding grade level or course placement, classroom summative evaluation, and program evaluation

fall into this category. Confidence in these decisions necessitates not only that more items be used than previously but that evidence to corroborate the test results be obtained. Teachers should incorporate all available information on a student's achievement in deciding which class he/she should be placed in or what letter grade he/she should receive in a specific course. When corroborative information is not concurrently available for placement decisions, confirmation of decision accuracy will depend upon the student's future performance. A decision error may be indicated by a student's poor performance in the subject into which he/she was placed. It then becomes rather inconvenient to transfer the student to another class at that point in time. The consequences of a class placement error areobviously more serious than those of a group placement error within a class. The error cannot be easily corrected. This serves to emphasize the importance of such decisions.

Importance of Objectives

Teachers generally consider some instructional and behavioral objectives more important than others. This is usually demonstrated by the amount of class time that is devoted to the different topics and activities in a given program. Basic skills may be stressed in a third-grade phonics program while more complex skills may be stressed in a twelfth-grade English composition course. The number of items corresponding to the objectives should proportionately reflect these relative weights.

The process by which objectives are weighted by item numbers is entirely judgmental. An instructional objective with three defining behavioral objectives, for example, may be measured with a total of seven items. This could be arrived at by analyzing each behavioral objective and assigning item numbers accordingly. If the objectives are arranged in order of increasing complexity, one, two, and four items may be written for the three objectives, respectively. This type of analysis is probably the most logical approach since each specific behavioral objective usually suggests how many as well as what type of items are most appro-

priate to measure it. The collection of items for a set of behavioral objectives constitutes the total number of items needed to assess the instructional objective.

When that procedure is followed for all instructional objectives, it systematically assures a balanced distribution of items both across and within the instructional objectives. The number of items frequently will differ from one objective to another. This variation can be expected if the items are to accurately represent the instructional emphases and the content domain.

Number of Objectives

One of the simplest guides to determining the total number of items per instructional objective is the number of behavioral objectives. It may provide an initial indication of how the items might be distributed prior to undertaking the weighting procedure described above. If there are four objectives at the knowledge level, for example, two items may be developed for each. When there is a long list of objectives, there are two options: either write one item for each objective or sample the objectives and write items for only the objectives that were sampled. The choice involved concerns the extent to which the final set of items represents the instructional objective.

One item per objective. Suppose there were 12 behavioral objectives related to one instructional objective. Each of the behavioral objectives defines a different aspect of the instructional objective. If all objectives were judged of equal importance, the task is simple. One item could be written for each objective. The 12 behavioral components of the instructional objective are, thereby, represented. This is preferable to randomly sampling six of the objectives and writing two items for each. Since there is no restriction on the number of items, the sampling is unnecessary.

Random sampling. When there is a restriction on the total number of items allotted for a given objective, sampling is essential. If, for example, only six items can be written for an instructional objective with 12 behavioral objectives,

six objectives should be randomly sampled and one item written for each. This sampling procedure assumes equal weighting for all of the objectives. It provides the most accurate measurement of the instructional objective under these conditions.

Stratified proportional allocation. Statements of behavioral objectives addressing different levels of cognition require a slight modification of the random sampling procedure. The list of 12 objectives would first be divided or stratified according to the levels of cognition. Five might be at the knowledge level, four at the comprehension level, and three at the application level. Instead of sampling six of the 12 objectives as before, the sampling should be proportionate to the number of objectives within each level or stratum. Three separate random samples of objectives might be drawn as follows: three out of five at the knowledge level, two out of four at the comprehension level, and one out of three at the application level. One item could then be written for each of the six objectives that were selected.

Stratified disproportional allocation. If particular emphasis is assigned to objectives in one or more strata, the weighting of those objectives can be incorporated into the sampling process. Based on the previous example, greater importance may be attached to the comprehension and application objectives than to the knowledge objectives. This can be reflected in the six objectives selected by sampling objectives disproportionately rather than proportionately to the numbers of objectives. More objectives would be selected at the two upper levels than at the lowest level regardless of the number of objectives at each level. The three random samples might be drawn as follows: one out of five at the knowledge level, three out of four at the comprehension level, and two out of three at the application level. This pattern almost reverses the proportionate sampling pattern. In taking into account the relative weights of the objectives, disproportionate random sampling will yield a set of behavioral objectives that is more representative of the instructional objective than the objectives chosen according

to strict proportionate sampling. It follows that the items constructed to measure the instructional objective should also be representative of that objective.

Practical Constraints

Item writing time. Once an approximate number of items has been determined for each instructional and behavioral objective based on the type of decisions, importance of objectives, and number of behavioral objectives, a review of the numbers should be conducted. The review should initially include an estimate of the time anticipated for writing the objective, short-answer, and/or essay items. The numbers often need to be reduced and redistributed in light of this estimate. There is a tendency to specify too many items for some objectives and not enough for others. When only a few hours are given to produce the items, a teacher should restrict the numbers and spend the time writing "quality" items rather than strive for a large collection of items at the expense of quality. Poor items are of little value in appraising what a student knows. The total number of items for a test measuring five instructional objectives should be between 25 and 50 items for placement and formative evaluation. The number may double for summative evaluation.

Administration time. A second consideration is the time available to administer the test. This is a function of the mode of administration. If the test is to be individually administered by the teacher on a one-to-one basis, the numbers should be minimized to practically accommodate the specific situation. Group administration offers more latitude. Although this administration is commonly restricted to the times of class periods or modules at the secondary level, there are three alternatives a teacher should consider to adapt the test administration to the class schedule: (1) administer the test in one sitting during a set time interval; (2) administer the test in sections in two or more sittings over several days; (3) administer one objective at a time prior to or after instruction on each objective. All three approaches are applicable to single class or an entire grade level. The first is the most restrictive in terms

of the number of items. Both the second and third approaches make it possible to include many more items per objective, if necessary. Administering the cluster of five to 10 items related to each objective as separate subtests will extend the administrations throughout the program. This may be particularly useful for formative evaluation decisions.

Students' characteristics. A third consideration is the characteristics of the students who will be tested. The students' grade level, ability level, and attention span may delimit the number of items that can be administered in a given period of time. If a student becomes bored, frustrated, or tired while taking a lengthy test, the quality and usefulness of the results will be diminished. The purpose of the test is to obtain a "true picture" of how much the student has learned. When his/her performance is distorted, so is this picture. The test should be structured according to the students' characteristics. It is sometimes more advantageous to break-up a comprehensive test into two or three parts and administer the parts at different times than to attempt a one-shot administration of the total test.

Application

A second-grade teacher used the foregoing guidelines to determine the numbers of items for her instructional and behavioral objectives in a unit on introductory dictionary skills. The objectives and item numbers are listed in Table 1. The process by which the four major factors were considered is explained below.

Insert Table 1 about here

Type of decisions. The test is intended to be used for placing students into groups based on their performance on the five instructional objectives. Non-masters of the first objective, for example, would be grouped for instruction on the skills related to that objective. This type of decision-making requires that

at least five items be written for each instructional objective.

Importance of objectives. The greatest emphasis in the unit was on locating entry words in the dictionary and syllabication. Alphabetizing, using guide words, and recognizing word meanings were judged about equal in importance but, overall, less important than the preceding objectives (2.0 and 3.0). An analysis of each behavioral objective was conducted to determine the appropriate number of items in light of the prespecified emphases. The illustrative item content provided with some of the objectives served to guide this process. It was decided, for example, that four items would be written to utilize the words cited in both objectives 3.2 and 3.3. Weights were also assigned to different behavioral objectives within a given instructional objective. The numbers of items specified for objectives 4.1 through 4.3 reflect this type of weighting. In order to assure the comprehensive measurement of all instructional objectives, a minimum of one item would be written for each defining behavioral objective. This coverage plus the weighting of the objectives according to their relative importance would yield a test that accurately represents the instruction and content of the dictionary skills unit.

Number of behavioral objectives. This factor was taken into account as part of the weighting process. The total number of items for each instructional objective reflects the number of behavioral objectives and, in some cases, the topics listed with the objectives. For example, the eight items indicated for objective 2.2 was arrived at by assigning two items to each of the dictionary headings for noun entries. Since there were no more than five behavioral objectives per instructional objective, sampling of objectives was unnecessary.

Practical constraints. The item numbers were reviewed in terms of the time needed to construct and administer the items. Due to the nature of the subject matter and the importance of the objectives, the numbers were not reduced. However, given the range of ability of the second-graders for whom the test was designed, it was decided to administer the 52 items over three sittings.

Eight items would be administered on the first day, 27 on the second day, and 17 on the third day. The first two sets of items were administered in paper-and-pencil format while the last set was to be administered orally by the teacher.

Summary

Four factors essential to determining the length of objectives-based criterion-referenced tests were delineated. The rationale and the procedures for applying the factors were described in terms of how teachers and evaluators develop the tests and use the results for placement, formative, summative, and program evaluation decisions. The specific guidelines that emerge should facilitate the construction of criterion-referenced tests at the classroom, school, system, and state levels.

References

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Berk, R. A. Item sampling from finite domains of written discourse. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Eignor, D. R., & Hambleton, R. K. The relationship of test length to criterion-referenced test reliability and validity. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.

Flaner, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Hambleton, R. K., Hutton, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), Problems in criterion-referenced measurement (CSE Monograph Series in Evaluation, No. 3). Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-275.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.

Wilcox, R. R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement, in press.

Table 1

Number of Items per Objective for Unit on Dictionary Skills

Instructional and Behavioral Objectives	No. of Items
The student will be able to:	
1.0 Alphabetize groups of words	8
1.1 Alphabetize words by first letter (e.g., ball, cat, hat, pit)	(2)
1.2 Alphabetize words by first and second letters (e.g., map, men, pin, rip, run)	(2)
1.3 Identify lists of words that are alphabetized from among those that are not	(4)
2.0 Locate entry words in the Word and Picture Dictionary	14
2.1 Define entry word	(1)
2.2 Locate noun entries under the headings: a. Words for People (e.g., baby, nurse, sailor) b. Words for Animals (e.g., duck, moose, lion) c. Words for Things (e.g., napkin, jeep, radio) d. Words for Places (e.g., Maryland, hospital, zoo)	(8)
2.3 Locate verb entries under the headings: a. Words for What We Do and Did (e.g., leap, mix, paint, peel) b. Words for What We Have Done (e.g., have run, have seen)	(3)
2.4 Locate adjective entries under the heading: Words that Tell What Kind or Color (e.g., brown, fat, quiet)	(2)
3.0 Syllabicate words	17
3.1 Define syllable	(1)
3.2 Identify the number of syllables (one or two) in spoken words (e.g., hat, line, blanket, insect)	(4)
3.3 Identify one and two syllable written words (e.g., grape, window, happy, paper)	(4)
3.4 State examples of one and two syllable words	(4)
3.5 Divide words into syllables using the dictionary (e.g., contest, continue)	(4)
4.0 Use guide words	7
4.1 State the purpose of guide words	(1)
4.2 Identify guide words on a dictionary page	(2)
4.3 Locate entries using guide words	(4)
5.0 Recognize word meanings	6
5.1 State how the dictionary indicates multiple meanings of words	(1)
5.2 Identify the most common meaning of words from among multiple meanings in the dictionary	(2)
5.3 Identify words that have been made from the root word (e.g., few)	(1)
5.4 Identify word meanings of pictures representing the words (e.g., a tree branch, jacks)	(2)